tmn 在 SemEval-2023 任务 9：使用 XLM-T、谷歌翻译和集成学习进行多语言推文亲密度检测

安娜·格拉兹科娃秋明大学
a.v.glazkova@utmn.ru

arXiv:2304.04054v1 [cs.CL] 8 Apr 2023

摘要

该论文介绍了一个基于Transformer的系统，用于 SemEval-2023任务9：多语言推文亲密度分析。该任务的目的是预测推文的亲密度，范围从 1（完全不亲密）到5（非常亲密）。竞赛的官方训练集包含六种语言（英语、西班牙语、意大利语、葡萄牙语、法语和中文）的推文。测试集除了包括上述六种语言外，还包含来自四种训练集中未出现的语言的外部数据（印地语、阿拉伯语、荷兰语和韩语）。我们提出了一种基于 XLM-T集成的解决方案。XLM-T是一种适配了Twitter 领域的多语言RoBERTa模型。为了提高未见语言上的性能，我们为每条推文补充了其英语翻译。我们探究了翻译数据对于微调中已见语言和未见语言的有效性，并评估了在基于Transformer 的模型中使用翻译数据的策略。我们的解决方案在排行榜上排名第4，在测试集上取得了 0.599 的总体皮尔逊相关系数。与所有 45个提交结果的平均分相比，所提出的系统将皮尔逊相关系数最高提升了 0.088。

1 引言

亲密度是语言的一个重要社会层面，有助于探索各种语境中现存的社会规范（Pei和Jurgens，2020年）。亲密度这一概念描述了个体如何在其感知到的相互依存、温暖以及亲自分享的意愿方面与其受话者建立关系（Perl- man 和 Fehr，1987年）。对语言中亲密度的自动评估能够为我们提供一幅更加清晰的社会互动和不同语言策略图景。

本文介绍了一个为 SemEval-2023 任务 9：多语言推文亲密度分析而开发的系统（Pei 等人，2022年）。该任务的目的是预测推文的亲密度。

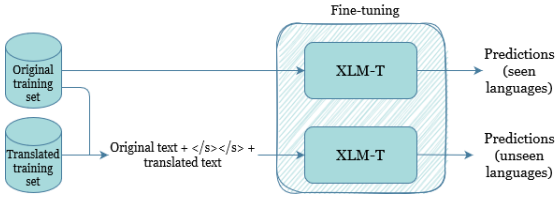


图1：我们方法的高层概述。

测试集包含训练集中出现的语言（已见语言）和未出现的语言（未见语言）的推文。该任务的主要挑战是对未见语言进行零样本亲密度预测的性能。

受近期TransformerMalmasi 等人，2022年；Tayyar Madabushi 等人，2022年），我们评估了 BERT（来自 Transformers 的双向编码器表示）（Devlin 等人，2019年）及其改进版本的性能。在我们的实验中，我们研究了翻译对提升零样本性能的影响。我们发现，与单独使用原始数据或翻译数据相比，同时使用原始文本和翻译文本能够提高模型在未见语言上的性能。

本文的组织结构如下。第 2节介绍任务以及任务组织者提供的数据集。我们的方法在第 3 节中介绍。在第 4 节中，我们描述了在竞赛的开发阶段所使用的实验设置。我们在第 5节中讨论实验结果。最后，第 6 节总结全文。

2 任务描述

该任务旨在预测十种语言推文的亲密度。组织者展示了一个训练集，其中包含六种语言（英语、西班牙语、意大利语、葡萄牙语、法语和中文）的推文，这些推文被标注了介于 1 和 5 之间的亲密度分数，其中 1 表示“完全不亲密”，5

语言	训练集	Test
数量（平均词元数）		
英语	1,587 (23.99)	396 (22.79)
西班牙语	1,592 (20.68)	399（20.05）
意大利语	1,532 (21.48)	384（20.71）
葡萄牙语	1,596（19.66）	398（20.69）
法语	1,588 (19.36)	393（20.43）
中文	1,596（29.97）	400（28.19）
印地语	-	280 (21.22)
阿拉伯语	-	407 (21.34)
荷兰语	-	413（19.6）
韩语	-	411 (19.37)

表 1：数据统计。词元数量使用 XLM-T 的分词器（Barbieri 等人，2022年）获得。

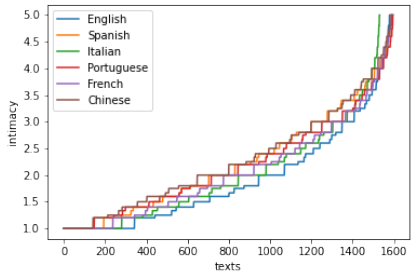


图 2：各语言亲密度分数的分布（训练集）。

表示“非常亲密”。模型在测试集上进行了评估，该测试集由给定的六种语言的推文以及未包含在训练数据中的四种未见语言（印地语、阿拉伯语、荷兰语和韩语）的推文组成。训练集中的条目总数量为 9,491。测试集的大小为 3,881。皮尔逊相关系数被用作评估指标。数据集统计信息见表 1。附录 A 中展示了英语已标注推文的示例。图 2 展示了训练集的亲密度分布。

3 系统概述

3.1 翻译后的推文

先前的研究表明，使用翻译后的数据集进行零样本学习的有效性（Schwenk 和 Douze，2017年；Eriguchi 等，2018；Ebrahimi 等，2022年）。我们使用公共翻译引擎谷歌翻译[1] 和 deep-translator Python 工具[2] 将训练集和测试集翻译成英语。我们比较了三种表示输入数据的策略：a）原始数据集，即将文本以其在原始数据集中的相同形式输入模型；b）翻译后的数据集，即将所有语言的所有文本替换为其英语译文；c）联合数据集，其中每个条目都包含由一串字符连接的原始文本和翻译文本 </s></s>。对于英语推文，原始推文与翻译后的推文一致。输入文本示例见表2。

3.2 模型

我们比较了多个预训练语言模型在推文亲密度预测任务上的表现：

• BERT[3] （Devlin等，2019年）：BERT多语-言基础模型（区分大小写），一种在104种语言上预训练的BERT多语言版本。

•XLM-R[4] （Conneau 等，2019年）：XLM-RoBERTa base，一个在包含100种语言的2.5TB经过筛选的CommonCrawl数据上预训练的模型。

• XLM-T[5] (Barbieri 等，2022年）：Twitter-XLM- Roberta 基础版，一种在超过 2 亿条推文上训练的多语言 RoBERTa 模型。

•RoBERTa[6] （Liu 等人，2019年）：RoBERTa 基础模型，是一种采用动态掩码进行预训练的 BERT 改进版本。由于 RoBERTa 是单语模型，我们仅使用翻译成英语的数据来使用它。

所有模型均使用基于 Transformers 的 Sim- ple Transformers[7] 库实现[8]（Wolf 等人，2020年）由 HuggingFace 提供。我们对每个模型进行了 3 个训练轮次的微调，使用批次大小 8、最大序列长度 128 和学习率 4e-5。

我们用于微调模型的代码以及所做的翻译已发布在 GitHub[9] 上。

[1] https://translate.google.com/[2] https://github.com/nidhaloff/deep-translator

[3] https://huggingface.co/bert-base-multilingual-cased[4] https://huggingface.co/xlm-roberta-base[5] https://huggingface.co/cardiffnlp/ t witter-xlm-roberta-base[6] https://huggingface.co/roberta-base[7] https://github.com/ThilinaRajapakse/ simpletransformers[8] https://github.com/huggingface/transformers[9] https://github.com/oldaandozerskaya/intimacy_tmn_ semeval23_task9

原文	已翻译	原文 +译文
我一点力量都没有了	我已经没有更多力量了	我已经没有任何力量了 </s></s> 我没有更多了力量
我的第一个赛季，也许是因此，我最喜欢的那个，好的 生日 #Reggina107 超文本传输协议	我的第一个赛季，也许是为了这个原因，我最喜欢的快乐 生日 #Reggina107 超文本传输协议	我的第一个赛季，也许正因为如此，我最喜欢的，生日快乐 #Reggina107 超文本传输协议 </s></s> 我的第一个赛季，也许对这个而言原因，我最喜欢的，生日快乐 #Reggina107 http
@user 当不.....时，这是正常的进展顺利	@user 事情不顺利时，这很正常 不顺利	@user 不顺利时很正常 </s></s> @user 事情不顺利时，这很正常
开学后更新会慢一点，快了 万粉了你们有什么想看的吗。 http	学校已经开学了，向上- 日期会慢一点，我是 将近 10,000 名粉丝，你有吗? 任何你想看的内容 超文本传输协议	开学了，会更新得慢一点了。快万粉了，你们有有什么想看的吗? http </s></s>学校有已启动，更新速度会稍慢一些，我正在将近 10,000 名粉丝，你有什么想要的东西吗? 想查看超文本传输协议

表2：输入表示的策略。

4 实验设置

在本节中，我们介绍竞赛开发阶段所采用的实验设置。在此阶段，组织者为参赛者提供了一个包含 9,491 个条目的训练集。我们使用固定的随机种子对所提供的数据进行打乱，并按 70:30 的比例将其划分为训练集和验证子集。因此，训练集和验证子集分别包含 6,643 个和 2,848 个条目。为了估计模型在未见语言上的表现，我们始终从训练集中排除一种语言。因此，模型在其余五种语言上进行微调，并在未见语言上评估。

为了评估开发阶段的结果，我们采用了两项指标。第一项是皮尔逊相关系数，是官方评估指标。第二项是均方误差（MSE），用于估算实际值与预测值之间距离平方的均值。

5 结果

5.1 开发阶段

我们比较了多个基于 Transformer 的预训练模型的表现，这些模型在开发阶段构建的训练子集上进行了微调。模型在未参与微调的验证子集上接受评估。验证子集中包含的所有语言也都出现在训练子集中。因此，验证子集中不包含未见语言。

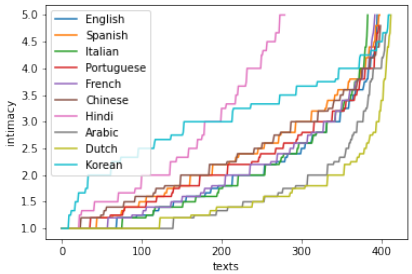
比较结果列于表 3 中。在均方误差和 Pearson相关系数r 方面，使用原始数据集进行微调的 XLM-T 取得了最佳结果。与 Pei 等人 （2022年）获得的分数类似，我们的结果显示了优越-

模型	*MSE*	*r*
原始数据		
BERT	0.562	0.18
XLM-R	0.499	0.517
XLM-T	**0.437**	**0.572**
翻译数据		
BERT	0.531	0.322
RoBERTa	0.495	0.431
XLM-R	0.488	0.422
XLM-T	0.494	0.534
原文 +译文		
BERT 0.545	0.353	
XLM-R	0.504	0.467
XLM-T	0.455	0.554

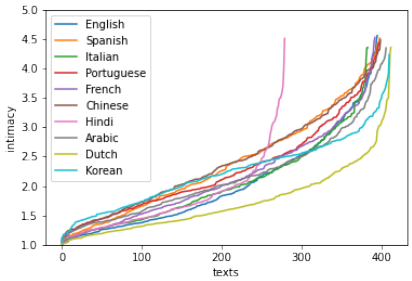
表3：开发阶段。在验证子集上的评估。

XLM-T 相较于 XLM-R 和 BERT在多语言推文亲密度预测任务中表现出优势。在翻译后的数据集上进行微调提升了 BERT 的结果；然而，与在原始数据集上进行微调所取得的结果相比，XLM-R 和 XLM-T的分数更低。联合使用原始数据和翻译数据在BERT模型中取得了最高的皮尔逊相关系数分数。总的来说，联合微调优于仅在翻译数据上进行微调，但未超过在原始文本上进行微调的模型的性能。就均方误差分数而言，结果并不总是与皮尔逊相关系数的值相对应。具体而言，XLM-R使用翻译后的数据集时取得了最佳的均方误差和最差的皮尔逊相关系数。

此外，我们评估了不同输入表示在零样本推文亲密度预测中的有效性。为此，我们逐一从原始训练集中排除各种语言。模型在其余语言上进行微调，然后在被排除的语言上进行评估



a) 金标准标签



b) 预测结果

图3：各语言的亲密度分数分布（测试集）。

。我们没有排除英语，因为在构建输入表示时，我们将英语用作辅助语言。结果如表4所示。第一列列出了训练集和验证子集所用的输入表示类型。例如，第一行展示了在原始数据上进行微调和评估的模型所获得的结果。

原始数据集的皮尔逊相关系数最低分数出现在中文（0.153）、法语（0.158）和意大利语（0.162）上。葡萄牙语的结果更具竞争力。对于西班牙语，我们取得了 0.579，高于表3中混合语言验证集上的分数（0.572）。对翻译成英语的数据进行微调，以及在原始数据集上微调后再用翻译数据进行验证，在大多数情况下都会导致结果变差。不过，使用翻译训练集时，中文的皮尔逊相关系数显著提升（+0.23）。对于在翻译文本上微调并在原始数据集上验证的模型，意大利语（皮尔逊相关系数提升+0.02）和法语（+0.069）也有所改善。意大利语和中文在翻译数据上的均方误差分数最佳。考虑到皮尔逊相关系数相对较低，可以推测模型预测的数值接近任意数据的平均值。就皮尔逊相关系数而言，对于五种语言中的三种（意大利语、法语和中文），联合使用原始数据和翻译数据进行微调的效果优于仅在原始数据集上微调。此外，原始数据的平均皮尔逊相关系数为 0.301，而联合使用原始文本和翻译文本时为 0.426。在我们的实验中，将翻译文本用作辅助数据后，平均皮尔逊相关系数提升了 0.125。

根据我们的实验，我们将由竞赛组织者提供的原始数据集上微调的 XLM-T，与同时利用原始数据和翻译文本进行联合微调的 XLM-T 结合起来。第一种模型用于对已见语言（英语、西班牙语、意大利语、葡萄牙语、法语和中文）进行预测，而第二种模型则应用于未见语言（印地语、荷兰语、韩语和阿拉伯语）。

5.2 总体性能

与我们之前关于推文分析的工作（Glazkova 等人，2021a，b）类似，我们在最终提交中应用了集成学习。我们采用了两部分方法，分别基于针对已见语言和未见语言的原始数据与联合数据进行微调的 XLM-T 模型集成（图1）。我们针对每种数据类型分别微调了七个模型，并对同一类型的所有模型取平均值。

表5展示了我们团队的官方成绩。所提出的模型在 SemEval-2023 Task 9 中排名第 4，测试集上的总体皮尔逊相关系数为 0.599。该结果比全部 45 支参赛队伍的平均分数高出 0.088。所提出的模型在数据集的所有语言上均取得高于平均水平的分数，并在已见语言和未见语言上均跻身前十。阿拉伯语（第 2 名）、印地语（第 5 名）、意大利语（第 6 名）和法语（第 9 名）的成绩最高。

5.3 误差分析

图 3 展示了官方测试集中的亲密度分数分布，以及使用我们的方法生成的预测结果分布。从图表可以看出，我们的模型预测的最大值低于官方数据集中的最大值。模型预测的最大值为 4.56，最小值为 0.99，而测试数据的取值范围为 1 至 5（含端点）。模型生成的分数

模型	未见语言									
	西班牙语		意大利语		葡萄牙语		法语		中文	
	MSE	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*
训练集：原始 验证集：原始	**0.414**	**0.579**	0.452	0.162	0.469	**0.452**	**0.435**	0.158	0.587	0.153
训练集：原始 验证集：已翻译	0.513	0.205	0.463	0.142	0.46	0.227	0.51	0.101	0.677	0.113
训练集：已翻译 验证集：已翻译	0.461	0.103	0.451	0.11	**0.455**	0.134	0.483	0.123	**0.541**	0.383
训练集：已翻译 验证集：原文	0.465	0.444	0.471	0.182	0.532	0.19	0.508	0.227	0.599	0.383
训练集、验证集： 原文 + 已翻译	0.42	0.565	**0.436**	**0.259**	0.462	0.439	0.459	**0.32**	0.577	**0.547**

表4：开发阶段。在未见语言上的评估。

分数	语言												
	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic	Seen	Unseen	Overall
r	0.717	0.74	0.684	0.734	0.708	0.721	0.242	0.639	0.361	0.662	0.637	0.357	0.599
Rank	11	10	12	6	9	15	5	9	17	2	9	9	4
平均 *r*	0.635	0.664	0.596	0.636	0.611	0.629	0.187	0.539	0.407	0.524	0.727	0.441	0.511

表5：官方结果。

更接近平均分数。官方分数的平均值为 2.13，标准差为 0.97。我们的预测结果平均值为 2.11，标准差为 0.74。

预测结果与真实测试标签之间的相关性如图4所示。总体而言，已见语言的结果可预期地更好。在所有语言中，印地语和韩语的相关性最低。

6 结论

在本研究中，我们探索了将预训练语言模型和翻译数据应用于多语言推文亲密度检测任务。我们证明，同时使用原始文本和翻译成英语的文本，可以提升 XLM-T 在未见语言上的零样本性能。未来可以开展关于零样本亲密度预测的进一步研究，以评估所用翻译的质量，并评估训练集和测试集语言之间的语言接近度。

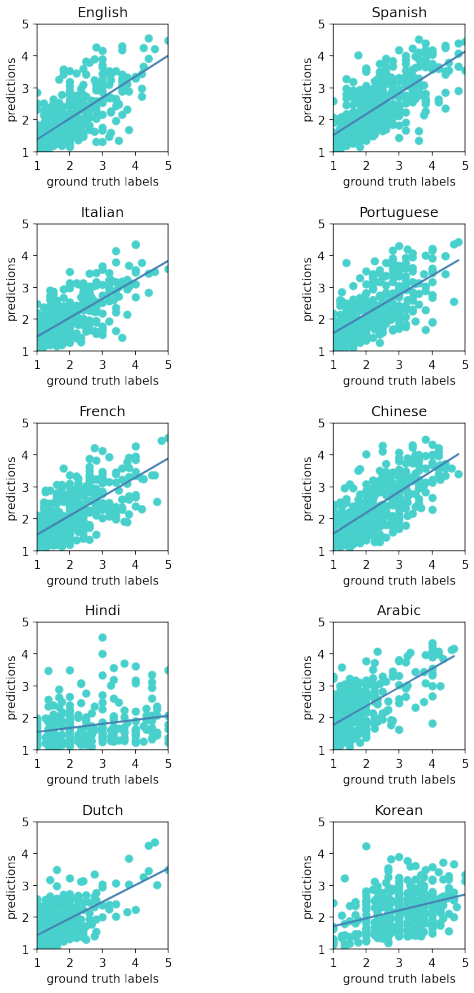


图4：预测结果与真实测试标签之间的相关性。

参考文献

Francesco Barbieri、Luis Espinosa Anke 和 Jose Camacho-Collados。2022年。XLM-T：用于情感分析及其他任务的 Twitter 多语言语言模型。载于第十三届语言资源与评测会议论文集，页码

258–266，马赛，法国。欧洲语言资源协会。

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettle- moyer 和 Veselin Stoyanov。2019年。大规模无监督跨语言表征学习。*CoRR*，abs/1911.02116。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深度双向 Transformers 预训练。载于*2019*年北美计算语言学协会分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，美国明尼阿波利斯。计算语言学协会。

Abteen Ebrahimi、Manuel Mager、Arturo Oncevay、Vishrav Chaudhary、Luis Chiruzzo、Angela Fan、John Ortega、Ricardo Ramos、Annette Rios、Ivan Vladimir Meza Ruiz、Gustavo Giménez-Lugo、Elisabeth Mager、Graham Neubig、Alexis Palmer、Rolando Coto-Solano、Thang Vu 和 Katharina Kann。2022年。AmericasNLI：在真正的低资源语言中评估预训练多语言模型的零样本自然语言理解能力。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第6279–6299页，爱尔兰都柏林。计算语言学协会。

AkikoEriguchi、Melvin Johnson、Orhan Firat、Hideto Kazawa 和 Wolfgang Macherey。2018 年。使用多语言神经机器翻译进行零样本跨语言分类。*arXiv* 预印本 *arXiv:1809.04686.pdf*。

Anna Glazkova、Maksim Glazkov 和 Timofey Tri-fonov。2021a。g2tmn 在 constraint@ AAAI2021：利用 CT-BERT 和集成学习进行 COVID-19 假新闻检测。载于 紧急情况下区域语言在线敌意帖子治理：第一届国际研讨会，*CONSTRAINT 2021*，与 *AAAI2021* 联合举办，*2021* 年 *2* 月 *8* 日，线上活动，修订版精选论文 *1*，第116–127 页。Springer。

Anna Glazkova、MichaelKadantsev 和 Maksim Glazkov。2021b。针对英语和马拉地语中的仇恨、攻击性及亵渎内容检测，对预训练变换器进行微调。载于 *FIRE 2021* 工作笔记——信息检索评估论坛，第 52–62 页。CEUR 研讨会论文集。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Shervin Malmasi、Anjie Fang、Besnik Fetahu、Sudipta Kar 和 Oleg Rokhlenko。2022年。SemEval-2022 任务11：多语言复杂命名实体识别（MultiCoNER）。载于第*16*届国际语义评测研讨会（*SemEval-2022*年）论文集，第1412–1437页，西雅图，美国。计算语言学协会。

Jiaxin Pei 和 David Jurgens。2020年。量化语言中的亲密度。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 5307–5326 页。

Jiaxin Pei、Vítor Silva、Maarten Bos、Yozon Liu、Leonardo Neves、David Jurgens 和 Francesco Barbieri。2022年。SemEval 2023 Task 9：多语言推文亲密度分析。*arXiv* 预印本 *arXiv:2210.01108*。

Daniel Perlman 和 Beverley Fehr。1987年。《亲密关系的发展》。

Holger Schwenk 和 Matthijs Douze。2017年。利用神经机器翻译学习联合多语言句子表示。载于自然语言处理表示学习第二届研讨会论文集，第157–167页，温哥华，加拿大。计算语言学协会。

Harish Tayyar Madabushi、Edward Gow-Smith、Marcos Garcia、Carolina Scarton、Marco Idiart 和 Aline Villavicencio。2022年。SemEval-2022任务2：多语言惯用性检测与句子嵌入。载于第*16*届语义评测国际研讨会（*SemEval-2022*）论文集，第107–121页，美国西雅图。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。2020年。Transformers：自然语言处理领域的先进技术。载于*2020*年自然语言处理实证方法会议：系统演示论文集，第38–45页，线上。计算语言学协会。

A 附录

推文	亲密度
这是一个漂亮的方程： 0+0-0-0+0=0	1.0
再试一次，蒂米……我会把你变成骨头- 餐食 超文本传输协议	1.4
@user @DoorDash 真可惜，你给了小费。 1.30 美元	2.25
@user @verge 我真不敢相信你竟然不重视家庭	3.0
我将永远属于你，直到岁月尽头 的时间。 #MewSuppasit @MSuppasit #My-CandyHeroxMew	4.0
朋友：你在对我笑什么？我：没什么，只是 我生命中的挚爱 我生命中的挚爱：超文本传输协议	4.4

表 6：数据集中的推文示例。